UNIVERSITI TEKNOLOGI MARA

EFFECTIVENESS OF SIMPLE TERMINOLOGICAL ONTOLOGY TO SUPPORT DOCUMENT RETRIEVAL IN A SPECIALISED DOMAIN

SEYED ABOLFAZLE MOOSAVIFAR

Thesis submitted in fulfillment
of the requirement for the degree of
Master of Science

Faculty of Computer and Mathematical Sciences

September 2014

AUTHOR'S DECLARATION

I declare that the work in this dissertation was carried out in accordance with the regulations of Universiti Teknologi MARA. It is original and is the results of my own work, unless otherwise indicated or acknowledged as referenced work. This topic has not been submitted to any other academic institution or non-academic institution for any degree or qualification.

I, hereby, acknowledge that I have been supplied with the Academic Rules and Regulation for Post Graduate, Universiti Teknologi MARA, regulating the conduct of my study and research.

Name of Student	:	Seyed Abolfazle Moosavifar
Student I.D. No.	:	2010230266
Programme	:	Master of Science
Faculty	:	Faculty of Computer and Mathematical Sciences
Thesis Title	:	Effectiveness of Simple Terminological Ontology to Support Document Retrieval in a Specialised Domain
Signature of Student	:	.....
Date	:	September 2014

ABSTRACT

The research investigated the proposition that a simple terminological ontology supported by general purpose lexical resources and aided by information retrieval and natural language processing techniques can effectively annotate and retrieve documents in a specialised knowledge domain. This is addressing the evidence from a recent survey, which reported that low satisfaction in the retrieval of documents in a personal collection. A common, but robust approach in this area is keyword-based retrieval. The weakness of keyword-based retrieval is its inability to 'understand' the meaning of the keywords (semantic). Ontology approach is introduced as a way to support semantic retrieval. However, there is a problem with the construction of the ontology by laymen, especially ontologies for specialised domain areas. Therefore, the use of simple terminological ontology (constructed based on intuitive understanding of the domain) is proposed in this research. The research objectives are structured to introduce new algorithms for ontology-based automatic annotation, retrieval and ranking of documents and to check on the reliability of WordNet to provide lexical support for the (simple terminological) ontology-based document retrieval. To achieve these objectives, the Boolean IR model was extended by incorporating four coefficients to adjust the term weights, namely to deal with the word significance and word coherence in multi-word terms, to consider the matching type (exact or synonym) and to factor the category weight when calculating the term weights. To find the retrieval effectiveness, the results of ontology-based retrieval was evaluated against the conventional retrieval, and validated against expert retrieval. The results of the ontology-based automatic annotation were evaluated against expert annotation. In addition, the reliability of using WordNet to provide lexical support was tested during the process of the annotation and retrieval. The research found synonyms from WordNet selected with the correct senses can help to improve the (simple terminological) ontology-based annotation and retrieval of documents in a specialised domain area. The research also found that (simple terminological) ontology-based retrieval that is support by selected synonyms from WordNet can recall all documents that are retrieved using keyword-based retrieval with reasonable precision. The evaluations of the retrieval by get help from expert domain also emphasized this result. The research result also indicated there are few common tags between the automatic and expert annotation. There were issues with the expert annotations; nonetheless, if we regard the expert annotation is paramount, then we suggest semi-automatic annotation of the documents in order to improve the result of ontology-based retrieval. Future researchers can use our research ideas (e.g. annotation and retrieval algorithms; assignment of weights to ontology terms) to make further progress in the field of semantic information retrieval. System designers can base our research findings (e.g. type of lexical support) to decide on methods for improving the retrieval in personal collection.

TABLE OF CONTENTS

AUTHOR'S DECLARATION ii

ABSTRACT iii

ACKNOWLEDGEMENT iv

TABLE OF CONTENTS viii

LIST OF TABLES x

LIST OF FIGURES xi

CHAPTER ONE INTRODUCTION

1.1 Background and Motivation 1

1.2 Problem Statement 5

1.3 Research Aim and Objectives 6

1.4 Research Scope 7

1.5 Research Significance 8

1.6 Thesis Organization 9

CHAPTER TWO LITERATURE REVIEW

2.1 Introduction 10

2.2 Archive Database 11

2.2.1 Text Database 12

2.2.2 Image Database 12

2.2.3 Other Databases 13

2.3 Ontology 13

2.3.1 Types of Ontology 15

2.3.2 Representation of Ontologies 26

2.3.3 Applications of Ontologies 28

2.4 Indexing 29

2.4.1 Manual Indexing 35

2.4.2 Automatic Indexing 36

2.4.3 Semi-Automatic Indexing 38

2.5 Natural Language Processing (NLP) Techniques 38
2.5.1 StopWord Removal 39
2.5.2 Morphological Normalization (MN) 40
2.5.3 Stemming 41
2.5.4 Part of Speech Tagging 42
2.6 Retrieval 44
2.6.1 Conventional IR Models 44
2.6.2 Semantic Information Retrieval 54
2.7 Evaluation of IR Systems 57
2.7.1 Recall and Precision in the IR System 57
2.7.2 F-Score 59
2.7.3 Mean Average Precision (MAP) 60
2.7.4 Mean Reciprocal Rank 62
2.8 Related Work 63
2.8.1 Conventional DR System 63
2.8.2 Ontology Based DR System 65
2.8.3 Ontology Based Retrieval in Specialized Domain 71
2.9 SUMMARY 73

CHAPTER THREE RESEARCH METHODOLOGY

3.1 Introduction 75
3.2 Database Model Design 75
3.3 Overall Framework For Conventional Document Retrieval 79
3.4 Overall Framework For Ontology Based Document Retrieval 79
3.5 Pre-Processing Ontology Terms 80
3.5.1 Domain Ontology Organization 81
3.5.2 Important Definitions 81
3.5.3 Sources of Synonyms and Stemmer 85
3.5.4 Decomposition of the Multi-Word Terms to Single Words 85
3.5.5 Matchable Multi-Word in the Document or User Query 86
3.5.6 Calculation of the Term-Document Weight 89
3.5.7 Calculation of the TF-IDF Weight 90
3.5.8 Calculation of a Single-Word Term Matching-Weight 90